Journal of Business Economics and Management
ISSN 1611-1699 / eISSN 2029-4433
2020 Volume 21 Issue 6: 1731–1751
https://doi.org/10.3846/jbem.2020.13194

COMPARATIVE MODELS IN CUSTOMER BASE ANALYSIS: PARAMETRIC MODEL AND OBSERVATION-DRIVEN MODEL

Shao-Ming XIE*

Department of Business Administration, National Taiwan University,
Taipei City, Taiwan, PRC

Received 20 May 2019; accepted 16 June 2020

Abstract. This study conducts a dynamic rolling comparison between the Pareto/NBD model (parametric model) and machine learning algorithms (observation-driven models) in customer base analysis, which the literature has not comprehensively investigated before. The aim is to find the comparative edge of these two approaches under customer base analysis and to define the implementation timing of these two paradigms. This research utilizes Pareto/NBD (Abe) as representative of Buy-Till-You-Die (BTYD) models in order to compete with machine learning algorithms and presents the following results. (1) The parametric model wins in transaction frequency prediction, whereas it loses in inactivity prediction. (2) The BTYD model outperforms machine learning in inactivity prediction when the customer base is active, performs better in an inactive customer base when competing with Poisson regression, and wins in a short-term active customer base when competing with a neural network algorithm in transaction frequency prediction. (3) The parametric model benefits more from a short calibration length and a long holdout/target period, which exhibit uncertainty. (4) The covariate effect helps Pareto/NBD (Abe) gain a better predictive result. These findings assist in defining the comparative edge and implementation timing of these two approaches and are useful for modeling and business decision making.

Keywords: BTYD, parametric model, Pareto/NBD model, observation-driven model, machine learning, customer base analysis, non-contractual setting.

JEL Classification: M31, C53.

Introduction

Relationship marketing emphasizes that a firm should maintain long-term relationships with its customers, because they help the firm derive more revenue (Benoit & Van den Poel, 2009; Gupta et al., 2006; Reinartz & Kumar, 2000). Unlike the situation when relationships between firms and customers are governed by a contract, non-contractual relationships commonly exist in many businesses, but they demand more attention from firms in order to manage their customer base. Marketing academics have developed a useful parametric model, the Pareto/

*Corresponding author. E-mail: D05741001@ntu.edu.tw

NBD model (Schmittlein et al., 1987) (Pareto/NBD (SMC), hereafter), to monitor a firm's customer base, and it has since become the golden standard for unearthing firm-customer relationships in non-contractual settings (Jerath et al., 2011) and is a high-efficiency model that needs only three frugal forms of information (Recency-Frequency-Calibration Length). Following this modeling framework, many researchers have promoted some useful variants, such as the BG/NBD model (Fader et al., 2005a), MBG/NBD model (Batislam et al., 2007), and periodic death opportunity (PDO, hereafter) model (Jerath et al., 2011). Among them, Abe (2009) provides a flexible alternative of the Pareto/NBD model (Pareto/NBD (Abe), hereafter) that incorporates richer customer characteristics as covariates and thus can utilize the increasing availability of customer transaction data with more information besides just recency, frequency, and calibration length.

Machine learning is commonly known as an observation-driven model and has permeated into every corner of many different industries (Ahmad et al., 2019; Coussement & De Bock, 2013; Smeureanu et al., 2013). It is able to detect patterns much easier and reuses uncovered patterns to predict future data (Murphy, 2012). Moreover, it provides numerous modeling candidates, like Logistic Regression (LG, hereafter), Poisson Regression (PR, hereafter), Decision Tree (DT, hereafter), Naïve Bayes (NB, hereafter), Support Vector Machine (SVM, hereafter), Random Forest (RF, hereafter), Neural Network Algorithm (NNA, hereafter), etc., which are easy to employ and could satisfy the needs of firms for discovering more valuable information from their customer base. Many studies have utilized machine learning in customer base analysis (Buckinx & Van den Poel, 2005; Kumar & Zymbler, 2019; Ngai et al., 2009), but BTYD models are relatively unknown by people. Both approaches provide solutions for customer base analysis, yet to the best of our knowledge, no previous research has conducted a comprehensive comparison between these two approaches under customer base analysis. Therefore, the first objective and contribution of this research are investigating the predictive edge between Pareto/NBD (Abe) and machine learning algorithms in customer base analysis.

Machine learning algorithms in this paper belong to supervised learning, which needs a certain time span to prepare the training label. Previous research studies have seldomly targeted the influence of the target/label span on the prediction results. Nie et al. (2011) define a customer as a churner who does not conduct any transaction during a 12-month period. Coussement and De Bock (2013) consider a gambler as a churner if he/she does not play during a 4-month period. Zhao et al. (2016) examine the sensitivity of predictive results to different label spans. Because these research studies do not explore the influence of the target/label span and holdout/prediction span on results and do not discuss the implementation timing between models, the second contribution of this research is to propose a labeling schema for modeling and to define the implementation scenario and timing of the two approaches.

The remainder of this paper runs as follows. The next section first reviews the BTYD models and explains the differences between the two approaches. Second, it introduces observation-driven models, including NNA, LG, PR, DT, RF, SVM, and NB, which are commonly used in marketing. This study then explores three real-world datasets and explains how the data are prepared for comparison. Next, the empirical results herein clarify the comparative edge between the parametric model and the observation-driven model. The study then conducts regression analysis to explore the effects of time span, data characteristics, and covariate effect among the comparative differences. Finally, this research concludes with discoveries, limitations, and future directions.

1. Model description and specification

Ngai et al. (2009) find that classification and association models have received the most research attraction, with customer retention analysis being the main application focus. Thus, DT, RF, SVM, and NB are also included in an inactivity comparison. NNA and LG are the main algorithms of machine learning for inactivity prediction. For transaction frequency prediction, NNA and PR are the chosen algorithms that can compete with Pareto/NBD (Abe).

1.1. Parametric model

1.1.1. Pareto/NBD model

Based on a customer's past transaction history, the Pareto/NBD model forecasts active status and purchase volume for a certain future period and builds upon two individual-level behavioral processes, the transaction process and the dropout process, which are depicted by Poisson distribution and exponential distribution. These two processes are assumed to be independent across customers, and heterogeneity among the customer base is modeled by two Gamma distributions. Following this framework, marketing scientists have accommodated this model to meet a wider array of application needs. Fader et al. (2005a) replace the dropout process with the Beta-Geometric paradigm (BG/NBD), which assumes that a dropout can occur immediately after a purchase. Fader et al. (2010) set up the BG/BB model that uses the Bernoulli-Beta paradigm to depict the transaction process, but it ignores the influence of previous transactions on present purchase behavior. Jerath et al. (2011) provide a variant, named the PDO model, that segregates discrete dropout opportunities from transaction time into calendar time. It allows customers to make a decision at a periodic length. These models use Maximum Likelihood Estimation (MLE, hereafter) to approximate the parameters, as it is an efficient method for estimating the Pareto/NBD model, but it encounters a severe problem due to numerous evaluations of the Gaussian Hypergeometric Function (Fader et al., 2005a; Ma & Liu, 2007).

Ma and Liu (2007) utilize Markov Chain Monte Carlo (MCMC, hereafter) for the estimation of Pareto/NBD (SMC) in order to solve the estimation burden of MLE, but they leave the derivations of the Pareto/NBD model intact (Singh et al., 2009). Abe (2009) takes advantage of the hierarchical Bayes framework (HB, hereafter) and MCMC and utilizes data augmentation (Tanner & Wong, 1987) to simplify the likelihood function when an unobservable lifetime and inactivity status are introduced as latent variables. In addition, he replaces the Gamma-Gamma prior distribution with the multivariate normal distribution to enable the correlation between the two processes and to introduce the covariate effect. His efforts improve computation efficiency and directly achieve useful individual-level estimations. In his empirical study, Pareto/NBD (Abe) with covariates performs better than that without covariates and demonstrates that recency-frequency could be conjuncted with a customer's characteristics and other behavior variables into customer base analysis. Platzer and Reutterer (2016) model the "clumpiness" idea raised in Zhang et al. (2014) in a more general timing pattern to capture regularity across customers and incorporate regularity into the Pareto/NBD model (named Pareto/GGG), but it cannot incorporate covariates in the case of the Gamma-Gamma-Gamma prior. Based on the above-mentioned improvements of Pareto/NBD (Abe) and its implementation advantages (Abe, 2009; Bernat, 2019; Korkmaz et al., 2013), this research employs Pareto/NBD (Abe) as the representative of BTYD models in order to compare with machine learning algorithms under customer base analysis.

1.1.2. Basic differences between the parametric model and observation-driven model

Before beginning the comparison, this study evaluates the parametric model versus the observation-driven model. Findings show that some basic differences between these two approaches may influence the acknowledgment of BTYD models in the business world.

HB could avoid the overfitting values through population distribution so as to structure dependency into the parameters (Dew & Ansari, 2018; Gelman et al., 2013). Pareto/NBD (Abe) is a parametric model – that is, each datapoint is used to fit its own likelihood, and then it maximizes the posterior function by MCMC. This means each datapoint has a series of parameter draws to achieve maximum a posteriori by marginalizing over all possible parameter choices. However, it may be too optimistic to use MCMC to maximize the posterior, due to the following reasons. (1) Irregularity transaction behavior or heterogeneity exists in the customer cohort, but with group characteristics. Individual estimation may dismiss valuable information from the group. (2) The aggregate information of transaction records may be insufficient enough to formulate an accurate distribution to depict these customers' true behavioral patterns via recency-frequency, thus leading to a greater risk of over-explanation.

Contrary to Pareto/NBD (Abe), the observation-driven model uses all datapoints to train the parameters on a universal aspect. As a learning algorithm, it learns the patterns of the data and not just one datapoint (Murphy, 2012; Witten et al., 2016). Hence, the weights of the observation-driven model capture the majority of characteristics in the customer cohort, which could be used to predict the out-of-sample. In addition, machine learning algorithms are much more flexible than Pareto/NBD (Abe) at adjusting their structure and meeting different kinds of data. Unlike the parametric model that makes several stringent assumptions on a limited number of variables, machine learning provides numerous innovative algorithms for marketers to handle a voluminous amount of data (Cui et al., 2006).

1.2. Observation-driven model

1.2.1. Neural Network Algorithm (NNA)

NNA is a network structure composed of Input Layer, Hidden Layer(s), and Output/Target Layer. A layer consists of neurons that control data transformation from the previous layer to the next layer. Between layers, neurons are connected so as to conduct the data stream from Input Layer to Output/Target Layer. This study adopts the fully connected neural network rather than other complex/deep NNAs, such as a Long Short-term Memory Network (LSTM) (Sifa et al., 2018) and Convolutional Neural Network (CNN) (Chen et al., 2018; Timoshenko & Hauser, 2019).

NNA can handle non-linear relationships between variables (West et al., 1997), and Fader et al. (2005b) find a non-linear relationship between recency-frequency and future transactions. In retention analysis, Ferreira et al. (2004) note that NNA dominates at inactivity prediction, and the best model has a structure with 15 hidden units. Sharma and Panigrahi (2011) also adopt a neural network-based approach for predicting inactivity, and the prediction accuracy of their proposed model exceeds 92%. The flexibility of the neural network is that it can be integrated with other models to generate a better prediction value. Hadden et al. (2007) join NNA with Genetic Algorithm, presenting empirical results that their model can powerfully predict customer inactivity.

To the best of our knowledge, there are scant pieces of research about transaction frequency prediction via NNA. Sifa et al. (2015) adopt the Poisson Regression Tree to predict the number of future purchases, by assuming a Poisson distribution for the purchases. However, the result is a binary tree that does not sufficiently utilize the meaning of "purchases". Sifa et al. (2018) focus on lifetime value prediction over a long period with 7 days of information, showing that the purchase amount and the number of previous purchases are the most informative features for predicting future customer lifetime values. They further find that transaction frequency is one of the most important features, but they are unable to provide a way to estimate the transaction frequency in the future.

NNA can easily be adapted to fit continuous variables when the loss function and activation function are replaced. One can utilize NNA with the Sigmoid function as the activation function and with Categorical Cross-Entropy as the loss function for fitting the active status. In transaction frequency, NNA is adopted with the tanh function in the hidden layer and Relu function in the output layer as the activation function and with Mean Square Error as the loss function. Thus, this paper uses trial-and-error to select the hidden nodes in the hidden layer and shows that NNA with 10 hidden nodes is able to generate the best predictive accuracy.

1.2.2. Logistic Regression (LG)

LG is a statistical technique that uses a logit transformation to map the outcome values from negative infinity to positive infinity, making it naturally suitable for inactivity prediction. Neslin et al. (2006) find that LG is commonly used by both academia and practitioners. In spite of Random Forest consistently performing the best, LG shows a similar prediction performance as both Random Forest and automatic relevance determination neural networks (Buckinx & Van den Poel, 2005). Nie et al. (2011) use credit card data of a Chinese bank to predict churners via Decision Tree and Logistic Regression, showing that LG performs better than Decision Tree in churn prediction.

1.2.3. Poisson Regression (PR)

One of the individual-level hypotheses of Pareto/NBD, the transaction process, follows a Poisson distribution. In transaction frequency analysis, the commonly used linear model for count data prediction is PR, which is a type of a generalized linear model where the response variable follows a Poisson distribution. Hence, this research considers PR for transaction frequency prediction. Coxe et al. (2009) summarize Poisson Regression and its variants in order to model count data. Some articles have also investigated the problems and the adaptations of PR at fitting count data (Gardner et al., 1995; Ver Hoef & Boveng, 2007). Trinh et al. (2014) propose the Poisson log-normal distribution, which replaces the Gamma distribution (prior distribution) with the log-normal distribution, for future purchase prediction, thus showing better performance toward buyer behavior than the negative binomial distribution.

1.2.4. Decision Tree (DT)

DT selects a variable's discernibility from high to low by information entropy. The commonly used evaluation methods are Information Gain, Information Gain Ratio, and Gini Index. This research utilizes DT with Information Gain. Hadiji et al. (2014) find that DT performs better than Neural Network Algorithm, Logistic Regression, and Naïve Bayes in terms of the weighted averaged F1-score. Hung et al. (2006) present that both Neural Network Algorithm and Decision Tree perform best at predicting churn, which helps a company know which customers will drop out. The results of DT are easily understandable and are able to achieve interpretable rules to instruct the prediction. Keramati et al. (2016) apply DT at churner prediction and extract the specific features of churners, thus helping bank managers to identify churners in the future.

1.2.5. Random Forest (RF)

RF is an ensemble learning algorithm that can solve the overfitting problem. It uses the feature of bagging to select those features that help achieve tree growth (Hastie et al., 2009). Burez and Van den Poel (2009) adopt the Weighted Random Forest in churn prediction, which performs significantly better than the Random Forest classifier. When denoting imbalanced data, the predictive class will be biased. Xie et al. (2009) thus incorporate both sampling techniques and cost-sensitive learning in RF to formulate an improved balanced random forest (IBRF). They find that the proposed algorithm performs better in churn prediction than other classifiers like the artificial neural networks, decision trees, and class-weighted core support vector machines (CWC-SVM).

1.2.6. Support Vector Machine (SVM)

For inactivity prediction, SVM targets to find a hyperplane that can segregate the classes. The hyperplane is supported by some representative datapoints to enlarge the gap between classes. Xia and Jin (2008) compare SVM with Decision Tree, Artificial Neural Network, Naïve Bayes, and Logistic Regression in the telecommunications industry, noting that SVM performs best in churn prediction. Coussement and Van den Poel (2008) combine SVM with a parameter-selection technique, which then executes better than Logistic Regression. However, the dataset has many features that are not linearly separable. The kernel function helps SVM to map the non-linear relationship into a high-dimensional space where the datapoints are linearly separable. Chen et al. (2012) formulate a hierarchical multiple kernel support vector machine (HMK-SVM) to compete with currently available classifiers, such as Decision, Boosting, Logistic Regression, etc. and discover that HMK-SVM exhibits outstanding performance under contractual and non-contractual settings.

1.2.7. Naïve Bayes (NB)

NB is a probabilistic model based on the Bayesian theorem that assumes attributes are conditionally independent. NB has been adopted by many research studies, but does not perform best among the classifiers (Buckinx et al., 2002; Saradhi & Palshikar, 2011; Vafeiadis et al., 2015). Huang et al. (2012) find that Naïve Bayes performs badly when facing a large number of features. They suggest using a dimension reduction technique, like Principal Component Analysis, to first transform features to a low dimension and then to employ Naïve Bayes in classification. This bad prediction performance may come from the independent assumption, which ignores the relationship/correlation between features.

2. Datasets and experimental set-up

2.1. Datasets

This study employs three different types of datasets – Mobile Game (GAME), Online Music Retailing (CDNOW), and Online Grocery Retailing (GROCERY) – in the comparison between Pareto/NBD (Abe) and machine learning. Table 1 reports data description for these three datasets.

Table 1. Data description

Key characteristic	GAME	CDNOW	GROCERY
Start date	2016-08-11	1997-01-01	2006-01-01
End date	2017-09-28	1998-06-30	2007-12-30
Type	Daily	Daily	Daily
Overall observations	189 339	14 658	10 483
Number of customers	5000	5000	1525
Sales			
Q25 Sales	20.00	14.49	
Median Sales	50.00	25.98	
Q75 Sales	270.00	44.10	
Average sales per customer	22 790.90	105.85	

The GAME dataset comes from a top-3 mobile game company in Taiwan. This dataset has a total of 5000 customers from 413 days of observations between 2016-08-11 and 2017-09-28. Marketing scientists have utilized the CDNOW dataset in many pieces of customer base prediction (e.g., Fader & Hardie, 2001; Romero et al., 2013; Wübben & Wangenheim, 2008; Zhang et al., 2014). To keep the same sample points as the GAME dataset, this research randomly samples 5000 customers. The observations are from a 545-day time window between 1997-01-01 and 1998-06-30. The GROCERY dataset is available from the BTYDplus package of R and is from an online retailer offering a broad range of grocery categories. There is no other information except customer ID and transaction date in this dataset. There are 10 483 transaction records made by 1525 customers during the observation period from 2006-01-01 to 2007-12-30.

The covariates are the in-App purchase in the GAME dataset and the expense in the CDNOW dataset. Both are named as "sales" in this research. From the statistics summary in Table 1, customers in the GAME dataset consume 22790.90 game coins on average, but this presents a positive skew. The distribution of individual transaction amount shows more asymmetry in the CDNOW dataset. Finally, the GROCERY dataset is naturally without covariates.

This research samples 50 customers of each dataset and visualizes their transaction records through timing patterns shown in Figure 1. The research finds that customers have unique transaction patterns in the different datasets: most customers are heavy users of GAME at the beginning but never come back after the last transaction; some customers do make repeat transactions across a long time period; most customers have a large inter-transaction time between transactions in CDNOW; and GROCERY has the most active customer base. Since differential transaction patterns exist in different datasets, the conclusion must be incorrect if this research only conveys one comparison between the parametric model and the observation-driven model. The next section introduces the dynamic analytical procedure for iterative comparison purposes.

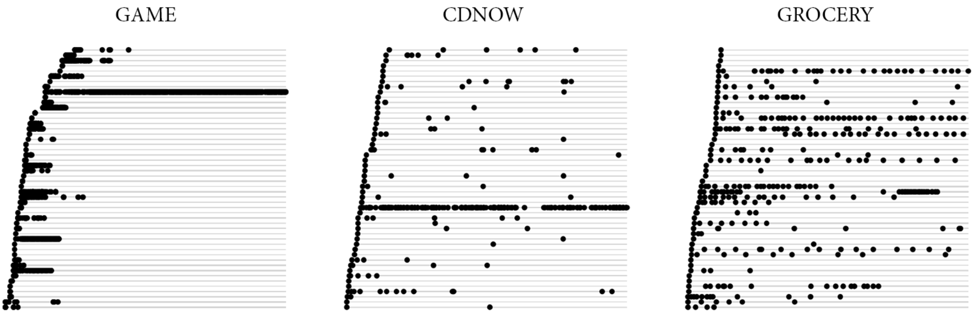


Figure 1. Fifty sampled customers' timing pattern in the three datasets

2.2. Estimation procedure

This research selects the supervised algorithms that need a target variable to train the algorithm's weights. As Figure 2 presents, they need the input data in (t_0 , $T - T^*$] and the target variable in ($T - T^*$, T], where t_0 is the first-ever transaction date, T is calibration date, and T^* is length of holdout/target period. For a fair comparison in testing between these two approaches, the information in the calibration period makes up the input variables – that is, recency-frequency and calibration length (covariates will be added if the dataset has more variables) are the input variables for the predictive comparison in (T , $T + T^*$].

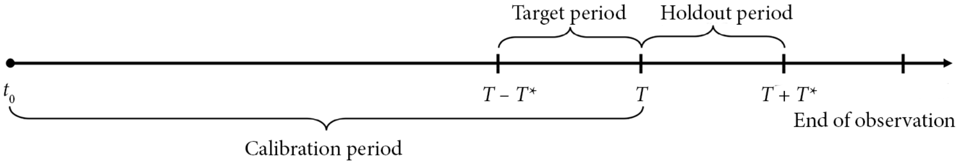


Figure 2. Illustration of data preparation for customer base prediction

Before moving on, this study must make a clarification about "Why does the Target Period Have an Equal Length as the Holdout Period?" First, the target variable in machine learning algorithms is extracted from the target period for training purposes, and thus the target period may affect testing accuracy. The equivalence of the target period and holdout period can eliminate the influence of the time span in testing. Second, the time span influences whether the customer is active in inactivity prediction. For example, there is higher inactivity potential in the seven-day target period than that in a one-day target period, because customers can flip the coin seven times rather than once. Third, it makes sense in the real business world when machine learning algorithms are utilized for predicting transactions of customers in a short-term or long-term period.

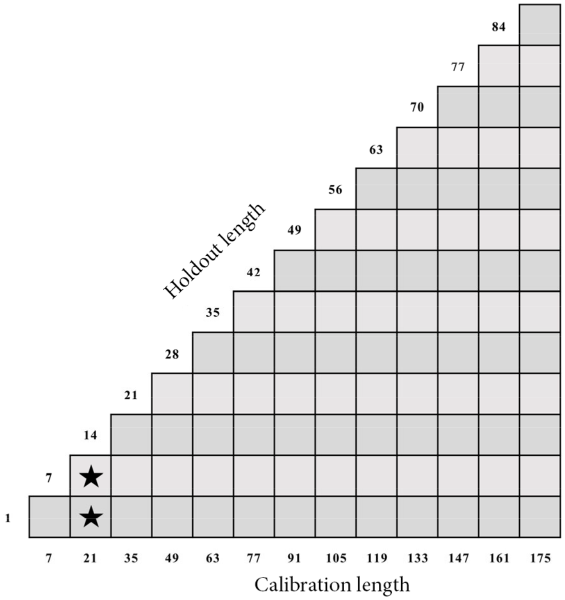


Figure 3. Illustration of the dynamic analytical procedure

The dynamic analytical procedure aims to clarify the influence of the time span on prediction accuracy. Figure 3 visualizes this dynamic comparison where the calibration length is on the horizontal axis, and the holdout/target period is on the vertical dimension. These two-dimensional scales split the comparison space into 91 combinations if the holdout/target length is constrained to be smaller than the calibration length. In addition, the holdout/target length is arranged weekly so that it satisfies managerial needs. In order to fully utilize the information of the dataset, each cell in Figure 3 is the basic unit where Pareto/NBD (Abe) competes with machine learning algorithms.

Figure 4 illustrates two specific examples to unravel the data preparation and the dynamic comparison when setting 21 days as the calibration period and when the target period and holdout period are set to 1 day and 7 days, respectively. These two comparisons are the two scenarios marked by a pentagram in Figure 3.

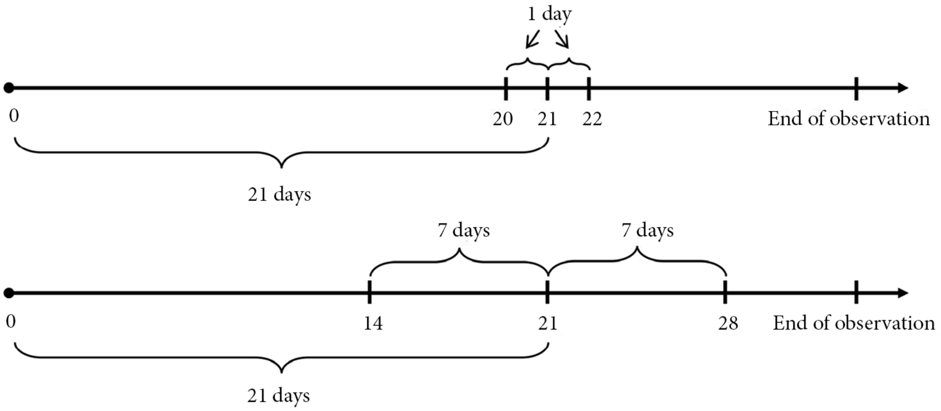


Figure 4. Specific interpretation of the analysis procedure

2.3. Evaluation index

This research uses two evaluation methods to assess the best model in each cell in Figure 3. This research adopts accuracy for inactivity prediction and Mean Absolute Error (MAE) for transaction frequency prediction.

1. Accuracy for Inactivity Evaluation

The confusion matrix is a commonly used evaluation method to summarize the performance of a classifier for the categorical classification task. Inactivity classification is a binary classification task, and the accuracy is then the ratio of the exact classified instances to the whole instances.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}}. \tag{1}$$

2. MAE for Transaction Frequency Evaluation

This research focuses on the average errors in transaction frequency prediction where all individual differences have equal weight. MAE is utilized for predictive ability comparison between the two approaches.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \tag{2}$$

where $\hat{y}_i$ is the fitted number of purchases, y_i is the actual future transaction frequency, and n denotes the number of customers.

3. Empirical results

This section investigates the empirical results of the three real-world datasets and compares the predictive edge between the two mentioned approaches. This study defines a "winner" in the combination as the model/algorithm having the best prediction accuracy. Table 2 and Table 4 show the number and percentage of different models, Figure 5 and Figure 6 show the specific winner in each cell if it has the best prediction accuracy, and Table 3 and Table 5 conclude the statistical testing results between the parametric model and observation-driven models, respectively.

3.1. Inactivity prediction

LG is generally the best model in the comparisons, because it has the highest winning numbers in the real-world datasets. Pareto/NBD (Abe) loses its predictive edge in all three real-world datasets, because it is unable to produce better accuracy than the observation-driven models. More importantly, the incorporated covariate seems inconducive for Pareto/NBD (Abe) to improve its prediction power, which implies that machine learning benefits more from the covariate.

Table 2. Number and percentage of different models winning at inactivity prediction

Dataset	DT	LG	NNA	PNBD	RF	SVM	NB
GAME	1 (1.10%)	25 (27.47%)	18 (19.78%)	17 (18.68%)	9 (9.89%)	21 (23.08%)	0 (0.00%)
GROCERY	11 (12.09%)	33 (36.26%)	12 (13.19%)	0 (0.00%)	18 (19.78%)	17 (18.68%)	0 (0.00%)
CDNOW	11 (12.09%)	22 (24.18%)	37 (40.66%)	0 (0.00%)	10 (10.99%)	11 (12.09%)	0 (0.00%)

Figure 5 shows that Pareto/NBD (Abe) is almost defeated by machine learning algorithms and could only protect its absolute prediction advantage for the long calibration length and holdout length in the GAME dataset. For this classification problem, different machine learning algorithms provide marked prediction accuracy even without any behavioral hypothesis like the parametric model. Moreover, NB does not show up in the best estimation results above, as it is a probabilistic model for point estimation with prior information from the training data. Thus, the rule-based model outperforms the probabilistic model, based on the behavioral hypothesis. LG, one of the simplest machine learning algorithms, is able to generate better predictive accuracy than Pareto/NBD (Abe).

Table 3. Paired *t*-test for inactivity prediction

Dataset	PNBD vs. BEST	PNBD vs. DT	PNBD vs. LG	PNBD vs. NNA	PNBD vs. RF	PNBD vs. SVM	PNBD vs. NB
GAME	–0.0056 (0.0180)	0.0031 (0.0897)	–0.0047 (0.0086)	–0.0038 (0.0040)	–0.0007 (0.7258)	0.0056 (0.0254)	0.0470 (0.0000)
GROCERY	–0.4241 (0.0000)	–0.4059 (0.0000)	–0.4211 (0.0000)	–0.3630 (0.0000)	–0.4086 (0.0000)	–0.4095 (0.0000)	–0.3720 (0.0000)
CDNOW	–0.1349 (0.0000)	–0.1321 (0.000)	–0.1336 (0.0000)	–0.1340 (0.0000)	–0.1327 (0.0000)	–0.1312 (0.0000)	–0.0659 (0.0000)

Note: BEST means the best machine learning algorithm that has the best prediction accuracy in each combination.

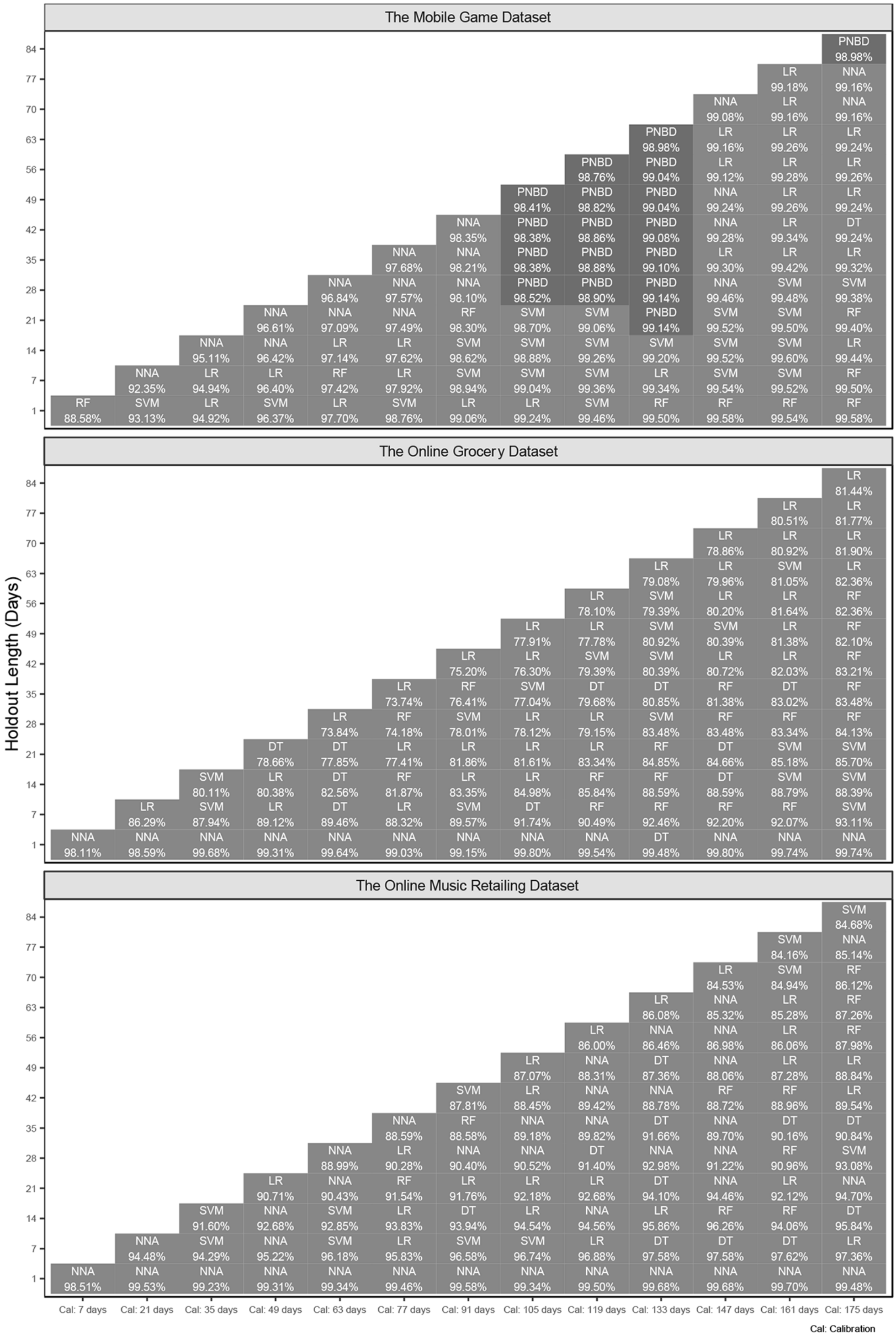


Figure 5. Best model at inactivity prediction

Aside from winner counting, this study uses the paired *t*-test to find the statistical significance between models. Table 3 shows that Pareto/NBD (Abe) has a significant insufficient predictive power over the machine learning algorithms in "PNBD vs. BEST". Pareto/NBD (Abe) is totally defeated in the CDNOW dataset and GROCERY dataset, but it has a better predictive edge over DT, SVM, and NB in the GAME dataset. This means that Pareto/NBD (Abe) loses in a general comparison, but wins in some one-to-one comparisons. Coupled with the timing pattern in Figure 1 and the mean of an individual in each dataset, Table 3 demonstrates that the machine learning algorithm wins by being more 13% or higher than Pareto/NBD (Abe) in infrequent datasets (CDNOW and GROCERY), but may lose in a frequent dataset (GAME).

3.2. Transaction frequency prediction

As an evaluation method for transaction frequency, measures the disagreement between the true transaction frequency and the predicted transaction frequency. This section compares Pareto/NBD (Abe) with NNA and PR at transaction frequency prediction.

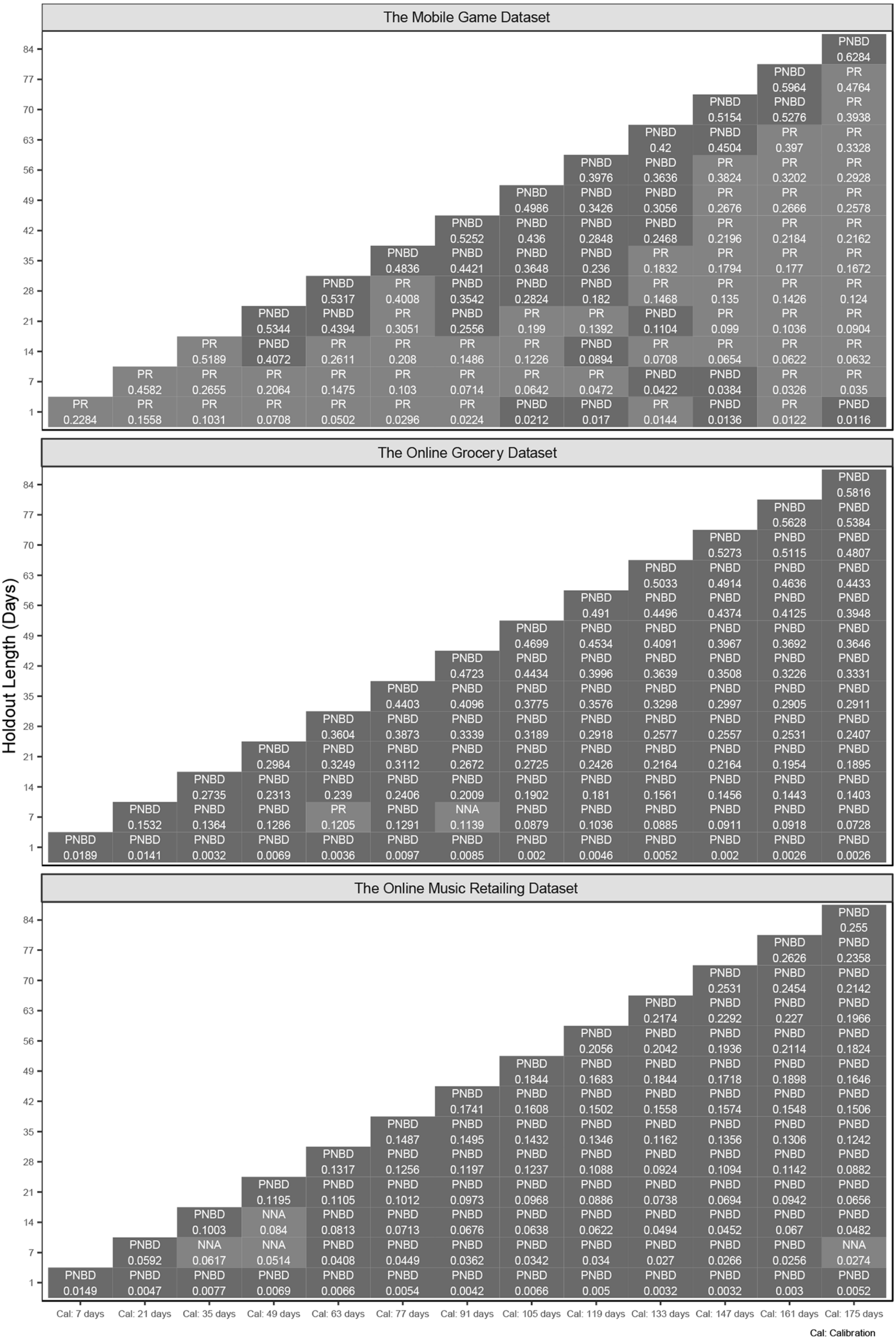
Table 4. Number and percentage of different models winning at transaction frequency prediction

Dataset	NNA	PNBD	PR
GAME	0 (0.00%)	35 (38.46%)	56 (61.54%)
GROCERY	1 (1.10%)	89 (97.80%)	1 (1.10%)
CDNOW	4 (4.40%)	87 (95.60%)	0 (0.00%)

NNA is completely beaten in this comparison. Pareto/NBD (Abe) dominates in this quantity's prediction, especially in the CDNOW dataset and the GROCERY dataset where it wins 87 scenarios and 89 scenarios, respectively. PR shows overwhelming advantages over NNA and is better than Pareto/NBD (Abe) in the GAME dataset. The included covariate has no covariate effect, because Pareto/NBD (Abe) has exactly the same performance in the GROCERY dataset (without covariate) as in the CDNOW dataset (with covariate). The model shows inconsistent performance in different datasets, which may be related to features of the customer base.

Figure 6 shows that the winning position of Pareto/NBD (Abe) is different in the three datasets. It dominates the CDNOW and GROCERY datasets where the calibration length and the target/holdout lengths have no influence on its performance, but unexpected performance appears in the GAME dataset. PR outperforms Pareto/NBD (Abe) and NNA 1) in the longest calibration length and the target/holdout length and 2) in the short calibration length and the target/holdout length, but Pareto/NBD (Abe) wins in the median calibration length and the long target/holdout length.

Different from inactivity prediction, Pareto/NBD (Abe) shows overwhelming predictive power over NNA and PR at transaction frequency forecasting. The customer-level behavioral hypothesis of Pareto/NBD (Abe) may contribute to the more accurate transaction frequency prediction.



The Mobile Game Dataset

Holdout Length (Days)	Cal: 7 days	Cal: 21 days	Cal: 35 days	Cal: 49 days	Cal: 63 days	Cal: 77 days	Cal: 91 days	Cal: 105 days	Cal: 119 days	Cal: 133 days	Cal: 147 days	Cal: 161 days	Cal: 175 days
84													PNBD 0.6284
77												PNBD 0.5964	PR 0.4764
70											PNBD 0.5154	PNBD 0.5276	PR 0.3938
63										PNBD 0.42	PNBD 0.4504	PR 0.397	PR 0.3328
56									PNBD 0.3976	PNBD 0.3636	PR 0.3824	PR 0.3202	PR 0.2928
49								PNBD 0.4986	PNBD 0.3426	PNBD 0.3056	PR 0.2676	PR 0.2666	PR 0.2578
42							PNBD 0.5252	PNBD 0.436	PNBD 0.2848	PNBD 0.2468	PR 0.2196	PR 0.2184	PR 0.2162
35						PNBD 0.4836	PNBD 0.4421	PNBD 0.3648	PNBD 0.236	PR 0.1832	PR 0.1794	PR 0.177	PR 0.1672
28					PNBD 0.5317	PR 0.4008	PNBD 0.3542	PNBD 0.2824	PNBD 0.182	PR 0.1468	PR 0.135	PR 0.1426	PR 0.124
21				PNBD 0.5344	PNBD 0.4394	PR 0.3051	PNBD 0.2556	PR 0.199	PR 0.1392	PNBD 0.1104	PR 0.099	PR 0.1036	PR 0.0904
14			PR 0.5189	PNBD 0.4072	PR 0.2611	PR 0.208	PR 0.1486	PR 0.1226	PNBD 0.0894	PR 0.0708	PR 0.0654	PR 0.0622	PR 0.0632
7		PR 0.4582	PR 0.2655	PR 0.2064	PR 0.1475	PR 0.103	PR 0.0714	PR 0.0642	PR 0.0472	PNBD 0.0422	PNBD 0.0384	PR 0.0326	PR 0.035
1	PR 0.2284	PR 0.1558	PR 0.1031	PR 0.0708	PR 0.0502	PR 0.0296	PR 0.0224	PNBD 0.0212	PNBD 0.017	PR 0.0144	PNBD 0.0136	PR 0.0122	PNBD 0.0116

The Online Grocery Dataset

Holdout Length (Days)	Cal: 7 days	Cal: 21 days	Cal: 35 days	Cal: 49 days	Cal: 63 days	Cal: 77 days	Cal: 91 days	Cal: 105 days	Cal: 119 days	Cal: 133 days	Cal: 147 days	Cal: 161 days	Cal: 175 days
84													PNBD 0.5816
77												PNBD 0.5628	PNBD 0.5384
70											PNBD 0.5273	PNBD 0.5115	PNBD 0.4807
63										PNBD 0.5033	PNBD 0.4914	PNBD 0.4636	PNBD 0.4433
56									PNBD 0.491	PNBD 0.4496	PNBD 0.4374	PNBD 0.4125	PNBD 0.3948
49								PNBD 0.4699	PNBD 0.4534	PNBD 0.4091	PNBD 0.3967	PNBD 0.3692	PNBD 0.3646
42							PNBD 0.4723	PNBD 0.4434	PNBD 0.3996	PNBD 0.3639	PNBD 0.3508	PNBD 0.3226	PNBD 0.3331
35						PNBD 0.4403	PNBD 0.4096	PNBD 0.3775	PNBD 0.3576	PNBD 0.3298	PNBD 0.2997	PNBD 0.2905	PNBD 0.2911
28					PNBD 0.3604	PNBD 0.3873	PNBD 0.3339	PNBD 0.3189	PNBD 0.2918	PNBD 0.2577	PNBD 0.2557	PNBD 0.2531	PNBD 0.2407
21				PNBD 0.2984	PNBD 0.3249	PNBD 0.3112	PNBD 0.2672	PNBD 0.2725	PNBD 0.2426	PNBD 0.2164	PNBD 0.2164	PNBD 0.1954	PNBD 0.1895
14			PNBD 0.2735	PNBD 0.2313	PNBD 0.239	PNBD 0.2406	PNBD 0.2009	PNBD 0.1902	PNBD 0.181	PNBD 0.1561	PNBD 0.1456	PNBD 0.1443	PNBD 0.1403
7		PNBD 0.1532	PNBD 0.1364	PNBD 0.1286	PR 0.1205	PNBD 0.1291	NNA 0.1139	PNBD 0.0879	PNBD 0.1036	PNBD 0.0885	PNBD 0.0911	PNBD 0.0918	PNBD 0.0728
1	PNBD 0.0189	PNBD 0.0141	PNBD 0.0032	PNBD 0.0069	PNBD 0.0036	PNBD 0.0097	PNBD 0.0085	PNBD 0.002	PNBD 0.0046	PNBD 0.0052	PNBD 0.002	PNBD 0.0026	PNBD 0.0026

The Online Music Retailing Dataset

Holdout Length (Days)	Cal: 7 days	Cal: 21 days	Cal: 35 days	Cal: 49 days	Cal: 63 days	Cal: 77 days	Cal: 91 days	Cal: 105 days	Cal: 119 days	Cal: 133 days	Cal: 147 days	Cal: 161 days	Cal: 175 days
84													PNBD 0.255
77												PNBD 0.2626	PNBD 0.2358
70											PNBD 0.2531	PNBD 0.2454	PNBD 0.2142
63										PNBD 0.2174	PNBD 0.2292	PNBD 0.227	PNBD 0.1966
56									PNBD 0.2056	PNBD 0.2042	PNBD 0.1936	PNBD 0.2114	PNBD 0.1824
49								PNBD 0.1844	PNBD 0.1683	PNBD 0.1844	PNBD 0.1718	PNBD 0.1898	PNBD 0.1646
42							PNBD 0.1741	PNBD 0.1608	PNBD 0.1502	PNBD 0.1558	PNBD 0.1574	PNBD 0.1548	PNBD 0.1506
35						PNBD 0.1487	PNBD 0.1495	PNBD 0.1432	PNBD 0.1346	PNBD 0.1162	PNBD 0.1356	PNBD 0.1306	PNBD 0.1242
28					PNBD 0.1317	PNBD 0.1256	PNBD 0.1197	PNBD 0.1237	PNBD 0.1088	PNBD 0.0924	PNBD 0.1094	PNBD 0.1142	PNBD 0.0882
21				PNBD 0.1195	PNBD 0.1105	PNBD 0.1012	PNBD 0.0973	PNBD 0.0968	PNBD 0.0886	PNBD 0.0738	PNBD 0.0694	PNBD 0.0942	PNBD 0.0656
14			PNBD 0.1003	NNA 0.084	PNBD 0.0813	PNBD 0.0713	PNBD 0.0676	PNBD 0.0638	PNBD 0.0622	PNBD 0.0494	PNBD 0.0452	PNBD 0.067	PNBD 0.0482
7		PNBD 0.0592	NNA 0.0617	NNA 0.0514	PNBD 0.0408	PNBD 0.0449	PNBD 0.0362	PNBD 0.0342	PNBD 0.034	PNBD 0.027	PNBD 0.0266	PNBD 0.0256	NNA 0.0274
1	PNBD 0.0149	PNBD 0.0047	PNBD 0.0077	PNBD 0.0069	PNBD 0.0066	PNBD 0.0054	PNBD 0.0042	PNBD 0.0066	PNBD 0.005	PNBD 0.0032	PNBD 0.0032	PNBD 0.003	PNBD 0.0052

Cal: Calibration

Figure 6. Best model at transaction frequency prediction

Table 5. Paired *t*-test in transaction frequency prediction

Dataset	PNBD vs. BEST	PNBD vs. NNA	PNBD vs. PR
GAME	0.0491 (0.0087)	0.3299 (0.000)	0.0547 (0.0072)
GROCERY	0.0453 (0.000)	0.0688 (0.0000)	0.1585 (0.1038)
CDNOW	0.0132 (0.0000)	0.0135 (0.0000)	5.5156 (0.0357)

Note: BEST means the best machine learning algorithm that has the best prediction accuracy in each combination.

3.3. Predictive difference decomposition

This research defines two dependent variables, $\text{Acc}_{\text{PNBD}} - \text{Acc}_{\text{competitor}}$ and $\text{MAE}_{\text{competitor}} - \text{MAE}_{\text{PNBD}}$, to clarify the advantage of Pareto/NBD (Abe) over machine learning algorithms. They have a positive value if Pareto/NBD (Abe) has a higher predictive accuracy over the selected machine learning algorithms. As with Table 1 data description and Figure 1 customer timing patterns, customers have different transaction patterns among different datasets. This study uses the average recency and average frequency of a customer base in the calibration period as the behavioral characteristics to examine their contribution for a predictive comparison. It includes the length of calibration period and holdout/target period to realize the effect of a time span and formulates the dummy variable to analyze the covariate effect.

Table 6. Regression results

Dependent variable	Inactivity		Transaction frequency	
	$\text{Acc}_{\text{PNBD}} - \text{Acc}_{\text{NNA}}$	$\text{Acc}_{\text{PNBD}} - \text{Acc}_{\text{LG}}$	$\text{MAE}_{\text{NNA}} - \text{MAE}_{\text{PNBD}}$	$\text{MAE}_{\text{PR}} - \text{MAE}_{\text{PNBD}}$
log(Avg (Recency))	0.185***	0.167***	–0.070*	2.181
	(0.027)	(0.023)	(0.038)	(2.795)
log(Avg (Frequency))	0.045***	0.045***	0.120***	–2.097***
	(0.008)	(0.007)	(0.011)	(0.799)
log(Time span for Calibration)	–0.170***	–0.155***	–0.062	–2.045
	(0.041)	(0.036)	(0.058)	(4.293)
log(Time span for Holdout/ Target)	0.058***	0.058***	0.074***	1.508**
	(0.007)	(0.006)	(0.010)	(0.728)
Covariate	0.420***	0.464***	0.006	5.065*
	(0.028)	(0.024)	(0.039)	(2.890)

End of Table 6

Dependent variable	Inactivity		Transaction frequency	
	$Acc_{PNBD} - Acc_{NNA}$	$Acc_{PNBD} - Acc_{LG}$	$MAE_{NNA} - MAE_{PNBD}$	$MAE_{PR} - MAE_{PNBD}$
Intercept	–0.284**	–0.362***	0.348**	–0.792
	(0.122)	(0.106)	(0.171)	(12.764)
R^2	0.710	0.788	0.432	0.048
Adjusted R^2	0.705	0.784	0.422	0.030

Notes: PNBD: Pareto/NBD (Abe); *$p < 0.01$, **$p < 0.001$, ***$p < 0.0001$.

If one considers significance, then the above regression result in Table 6 indicates the following.

For inactivity prediction, Pareto/NBD (Abe) performs better than NNA or LG when the calibration period is short and the holdout/target period is long. Pareto/NBD (Abe) excels in a long-term active customer base, which is characterized by a relative larger average transaction frequency and recency.

Comparing NNA at transaction frequency prediction, Pareto/NBD (Abe) performs better at long-term prediction for a short-term active customer base, which is characterized by a large average transaction frequency and small average recency. When comparing to PR, Pareto/NBD (Abe) dominates in long-term prediction for an inactive customer base. Just like inactivity prediction, Pareto/NBD (Abe) can sustain more severe uncertainty for a long target/holdout period.

The covariate has an insignificant effect in a comparison between Pareto/NBD (Abe) and NNA on transaction frequency prediction, while Pareto/NBD (Abe) benefits more when competing with PR. Conversely, the covariate has a significant effect on inactivity prediction. More customer purchasing information helps Pareto/NBD (Abe) gain higher predictive accuracy than NNA and LG.

3.4. Discussions

In the visualized results and the absolute winner counting, the findings show that Pareto/NBD (Abe) cannot compete with machine learning at inactivity prediction, but nearly rules over all the transaction frequency prediction scenarios. The results of the paired *t*-test indicate that the parametric model has a dominant edge in transaction frequency prediction, but is almost defeated at inactivity prediction even with some winning in the GAME dataset. Therefore, the relationship between machine learning and Pareto/NBD (Abe) is stable over different combinations of calibration period and holdout/target period.

The regression results in Table 6 demonstrate that Pareto/NBD (Abe) is an expert at inactivity prediction when the customer base is long-term active. For transaction frequency prediction, Pareto/NBD (Abe) wins for an active customer base when competing with NNA and for an inactive customer base when competing with PR. Moreover, Pareto/NBD (Abe) has a dominant advantage in a short calibration length and long holdout/target length, when the training dataset comprises severe uncertainty. Different from inactivity prediction, Pareto/NBD (Abe) can only benefit from the covariate when comparing with PR. Additionally, the covariate effect does not exist in the absolute winning comparison between approaches, but more purchasing information helps Pareto/NBD (Abe) gain a higher predictive edge over machine learning.

Conclusions

Following the achieved results in this paper's empirical analysis, winner counting and the paired *t*-test in general indicate that the parametric model wins at transaction frequency prediction and that the observation-based model dominates for inactivity prediction. Pareto/NBD (Abe) has a predominant advantage under a short calibration period and a long target/holdout period where machine learning performs badly. Thus, this research examines the influence of data characteristics on a model's comparative edge given the average recency and average frequency of a customer base. Findings show that Pareto/NBD (Abe) wins at inactivity prediction when the customer base is active. Given an inactive customer base, Pareto/NBD (Abe) outperforms PR at transaction frequency prediction, but loses its predictive edge when competing with NNA. Furthermore, Pareto/NBD (Abe) benefits more from the covariate effect, which helps to narrow the predictive difference between two approaches.

The empirical results define the comparative edge of these two approaches and thus offer some managerial implications. First, managers and practitioners can select a specific modeling approach to obtain valuable information from the data. This study suggests that the observation-driven model may be a replacement for the parametric model for inactivity prediction, but the empirical results show that the latter has a better fit than the former at transaction frequency prediction. This provides evidence why the classification has received the most research attraction and that customer retention analysis is the main application focus. Hence, managers and practitioners can utilize machine learning for inactivity prediction and the BTYD model for transaction frequency prediction. Besides the inactivity prediction, managers have the ability to make a better inventory management if they combine customer image and basket analysis with the transaction frequency predicted by the parametric model. In other words, managers can make inventory management at individual level when they know what the customer looks like by customer image, what his/her most favorite goods or services are by basket analysis, and the times that the customer will revisits.

Second, the results of regression analysis help to clarify the implementation timing of the two approaches. For example, the parametric model has high tolerance for uncertainty in the short calibration length and the long holdout/target length. Practitioners may benefit from this discovery by obtaining a more accurate prediction when facing a barren dataset. Furthermore, the covariate is helpful at distinguishing the implementation timing and comparative edge for both inactivity prediction and transaction frequency prediction, which means that more customer purchasing information will make the model's prediction better. Hence, the covariate helps narrow the predictive difference between the two approaches, and Pareto/NBD (Abe) can gain better prediction results. Hence, business analysts depend on the richness of data to apply the right model at right time, then to support business managers in business monitoring and decision making.

Third, this research provides a label preparation schema that eliminates the influence and noise of the target span and holdout period. This schema differs from previous studies that only convey limited comparisons and do not fully utilize information at different calibration lengths. Furthermore, managers and practitioners can harness the labeling schema in this research to gain a more reasonable and accurate predictive model under different calibration lengths and holdout periods. Besides the technical aspect of the proposed labelling schema, it empowers managers to connect their business projects with business intelligence (BI) from the decision making aspect. Hence, they will know how many resources they can coordinate and allocate to their projects in a reasonable holdout period.

Aside from these plentiful results and benefits, three future directions offer targets of interest for follow-up research. The main limitation of this present study is that the time-invariant variable is absent, which may benefit the comparison if datasets include various characteristics on customers. In addition, only one covariate is included herein, which may not be able to fully employ the covariate effect into the decomposition of the comparative edge. Hence, it would be worth it to conduct a further study if more time-invariant or time-variant variables are available. Moreover, the only adopted BTYD model is Pareto/NBD (Abe), which may result in an unfair and unbalanced comparison. For a non-covariate comparison, several BTYD models can be made available for competing with machine learning under customer base analysis, and future research should thus be able to obtain more robust and comprehensive results.

Different datasets in different calibration periods have different customer base characteristics. This research uses the average recency and average frequency of a customer base to explain the predictive differences therein, but there are some readily available concepts that can be made substitutes, such as the previously mentioned “clumpiness” or “regularity”. Additionally, this research utilizes the standard version of machine learning algorithms. Other deep learning structures can satisfy the comparison needs, such as LSTM and CNN. These network structures should meet researchers’ desire to obtain more useful information for comparison. These continuous models are helpful at exploring sequential data and may obtain more accurate predictions than machine learning. Lastly, future research can employ an ensemble machine learning algorithm and deep learning structure to explore customer data and purchasing data simultaneously.

References

Abe, M. (2009). “Counting your customers” one by one: A hierarchical Bayes extension to the Pareto/NBD model. *Marketing Science*, *28*(3), 541–553. https://doi.org/10.1287/mksc.1090.0502

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1), 28. https://doi.org/10.1186/s40537-019-0191-6

Batislam, E. P., Denizel, M., & Filiztekin, A. (2007). Empirical validation and comparison of models for customer base analysis. *International Journal of Research in Marketing*, *24*(3), 201–209. https://doi.org/10.1016/j.ijresmar.2006.12.005

Benoit, D. F., & Van den Poel, D. (2009). Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services. *Expert Systems with Applications*, *36*(7), 10475–10484. https://doi.org/10.1016/j.eswa.2009.01.031

Bernat, J. R. (2019). *Modelling customer lifetime value in a continuous, non-contractual time setting.* http://hdl.handle.net/2105/45923

Buckinx, W., Baesens, B., Van den Poel, D., Van Kenhove, P., & Vanthienen, J. (2002). Using machine learning techniques to predict defection of top clients. *WIT Transactions on Information Communication Technologies*, *28*.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, *164*(1), 252–268. https://doi.org/10.1016/j.ejor.2003.12.010

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*(3), 4626–4636. https://doi.org/10.1016/j.eswa.2008.05.027

Chen, P. P., Guitart, A., del Río, A. F., & Periáñez, Á. (2018). Customer lifetime value in video games using deep learning and parametric models. In *2018 IEEE International Conference on Big Data (Big Data),* (pp. 2134–2140). IEEE. https://doi.org/10.1109/BigData.2018.8622151

Chen, Z. Y., Fan, Z. P., & Sun, M. H. (2012). A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of Operational Research*, *223*(2), 461–472. https://doi.org/10.1016/j.ejor.2012.06.040

Coussement, K., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *Journal of Business Research*, *66*(9), 1629–1636. https://doi.org/10.1016/j.jbusres.2012.12.008

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, *34*(1), 313–327. https://doi.org/10.1016/j.eswa.2006.09.038

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment*, *91*(2), 121–136. https://doi.org/10.1080/00223890802634175

Cui, G., Wong, M. L., & Lui, H.-K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science*, *52*(4), 597–612. https://doi.org/10.1287/mnsc.1060.0514

Dew, R., & Ansari, A. (2018). Bayesian nonparametric customer base analysis with model-based visualizations. *Marketing Science*, *37*(2), 216–235. https://doi.org/10.1287/mksc.2017.1050

Fader, P. S., & Hardie, B. G. (2001). Forecasting repeat sales at CDNOW: A case study. *Interfaces*, *31*(3_suppl.), S94-S107. https://doi.org/10.1287/inte.31.4.94.9683

Fader, P. S., Hardie, B. G., & Lee, K. L. (2005a). "Counting your customers" the easy way: An alternative to the Pareto/NBD model. *Marketing Science*, *24*(2), 275–284. https://doi.org/10.1287/mksc.1040.0098

Fader, P. S., Hardie, B. G., & Lee, K. L. (2005b). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, *42*(4), 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415

Fader, P. S., Hardie, B. G., & Shang, J. (2010). Customer-base analysis in a discrete-time noncontractual setting. *Marketing Science*, *29*(6), 1086–1108. https://doi.org/10.1287/mksc.1100.0580

Ferreira, J., Vellasco, M. M., Pacheco, M. A. C., Carlos, R., & Barbosa, H. (2004). *Data mining techniques on the evaluation of wireless churn* [Conference presentation]. European Symposium on Artificial Neural Networks, Bruges, Belgium.

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, *118*(3), 392. https://doi.org/10.1037/0033-2909.118.3.392

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Taylor & Francis. https://doi.org/10.1201/b16018

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., & Sriram, S. (2006). Modeling customer lifetime value. *Journal of Service Research*, *9*(2), 139–155. https://doi.org/10.1177/1094670506293810

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, *34*(10), 2902–2917. https://doi.org/10.1016/j.cor.2005.11.007

Hadiji, F., Sifa, R., Drachen, A., Thurau, C., Kersting, K., & Bauckhage, C. (2014). Predicting player churn in the wild. In *2014 IEEE Conference on Computational Intelligence and Games* (pp.1–8). IEEE. https://doi.org/10.1109/CIG.2014.6932876

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2^{nd} ed.). Springer Science & Business Media.

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, *39*(1), 1414–1425. https://doi.org/10.1016/j.eswa.2011.08.024

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, *31*(3), 515–524. https://doi.org/10.1016/j.eswa.2005.09.080

Jerath, K., Fader, P. S., & Hardie, B. G. (2011). New perspectives on customer "death" using a generalization of the Pareto/NBD model. *Marketing Science*, *30*(5), 866–880. https://doi.org/10.1287/mksc.1110.0654

Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, *2*(1), 10. https://doi.org/10.1186/s40854-016-0029-6

Korkmaz, E., Kuik, R., & Fok, D. (2013). *"Counting Your Customers": When will they buy next? An empirical validation of probabilistic customer base analysis models based on purchase timing* (ERIM Report Series Research in Management, ERS-2013-2001-LIS). Erasmus Research Institute of Management. http://hdl.handle.net/1765/38235

Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. *Journal of Big Data*, *6*(1), 62. https://doi.org/10.1186/s40537-019-0224-1

Ma, S.-H., & Liu, J.-L. (2007). The MCMC approach for solving the Pareto/NBD model and possible extensions. In *Third International Conference on Natural Computation (ICNC 2007).* (pp. 505–512). IEEE. https://doi.org/10.1109/ICNC.2007.728

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, *43*(2), 204–211. https://doi.org/10.1509/jmkr.43.2.204

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, *38*(12), 15273–15285. https://doi.org/10.1016/j.eswa.2011.06.028

Platzer, M., & Reutterer, T. (2016). Ticking away the moments: Timing regularity helps to better predict customer activity. *Marketing Science*, *35*(5), 779–799. https://doi.org/10.1287/mksc.2015.0963

Reinartz, W. J., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing*, *64*(4), 17–35. https://doi.org/10.1509/jmkg.64.4.17.18077

Romero, J., Van der Lans, R., & Wierenga, B. (2013). A partially hidden Markov model of customer dynamics for CLV measurement. *Journal of Interactive Marketing*, *27*(3), 185–208. https://doi.org/10.1016/j.intmar.2013.04.003

Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. *Expert Systems with Applications*, *38*(3), 1999–2006. https://doi.org/10.1016/j.eswa.2010.07.134

Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers: Who-are they and what will they do next? *Management Science*, *33*(1), 1–24. https://doi.org/10.1287/mnsc.33.1.1

Sharma, A., & Panigrahi, D. (2011). A neural network based approach for predicting customer churn in cellular network services. *International Journal of Computer Applications*, *27*(11), 26–31. https://doi.org/10.5120/3344-4605

Sifa, R., Hadiji, F., Runge, J., Drachen, A., Kersting, K., & Bauckhage, C. (2015). *Predicting purchase decisions in mobile free-to-play games* [Conference presentation]. Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference.

Sifa, R., Runge, J., Bauckhage, C., & Klapper, D. (2018). Customer lifetime value prediction in noncontractual freemium settings: Chasing high-value users using deep neural networks and SMOTE. In *Proceedings of the 51st Hawaii International Conference on System Sciences.* https://doi.org/10.24251/HICSS.2018.115

Singh, S. S., Borle, S., & Jain, D. C. (2009). A generalized framework for estimating customer lifetime value when customer lifetimes are not observed. *Quantitative Marketing and Economics*, *7*(2), 181–205. https://doi.org/10.1007/s11129-009-9065-0

Smeureanu, I., Ruxanda, G., & Badea, L. M. (2013). Customer segmentation in private banking sector using machine learning techniques. *Journal of Business Economics and Management*, *14*(5), 923–939. https://doi.org/10.3846/16111699.2012.749807

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528–540. https://doi.org/10.1080/01621459.1987.10478458

Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*, *38*(1), 1–20. https://doi.org/10.1287/mksc.2018.1123

Trinh, G., Rungie, C., Wright, M., Driesener, C., & Dawes, J. (2014). Predicting future purchases with the Poisson log-normal model. *Marketing Letters*, *25*(2), 219–234. https://doi.org/10.1007/s11002-013-9254-1

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1–9. https://doi.org/10.1016/j.simpat.2015.03.003

Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, *88*(11), 2766–2772. https://doi.org/10.1890/07-0043.1

West, P. M., Brockett, P. L., & Golden, L. L. (1997). A comparative analysis of neural networks and statistical methods for predicting consumer choice. *Marketing Science*, *16*(4), 370–391. https://doi.org/10.1287/mksc.16.4.370

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.

Wübben, M., & Wangenheim, F. v. (2008). Instant customer base analysis: Managerial heuristics often "get it right". *Journal of Marketing*, *72*(3), 82–93. https://doi.org/10.1509/jmkg.72.3.082

Xia, G. E., & Jin, W. D. (2008). Model of customer churn prediction on support vector machine. *Systems Engineering – Theory & Practice*, *28*(1), 71–77. https://doi.org/10.1016/S1874-8651(09)60003-X

Xie, Y. Y., Li, X., Ngai, E., & Ying, W. Y. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, *36*(3), 5445–5449. https://doi.org/10.1016/j.eswa.2008.06.121

Zhang, Y., Bradlow, E. T., & Small, D. S. (2014). Predicting customer value using clumpiness: From RFM to RFMC. *Marketing Science*, *34*(2), 195–208. https://doi.org/10.1287/mksc.2014.0873

Zhao, Y., Yao, L., & Zhang, Y. (2016). Purchase prediction using Tmall-specific features. *Concurrency Computation: Practice Experience*, *28*(14), 3879–3894. https://doi.org/10.1002/cpe.3720